Massively parallel mapping of substrate cleavage sites defines dipeptidyl peptidase four subsite cooperativity

Rajani Kanth Gudipati[1,2,*], Dimos Gaidatzis[1,3,*], Jan Seebacher[1], Sandra Muehlhaeusser[1], Georg Kempf[1], Simone Cavadini[1], Daniel Hess[1], Charlotte Soneson[1,3], and Helge Großhans[1,4,5]

[1] Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, Basel 4058, Switzerland.
[2] Centre for Advanced Technology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 10, 61-614 Poznań, Poland.
[3] SIB Swiss Institute of Bioinformatics, Basel, Switzerland.
[4] Faculty of Natural Sciences, University of Basel, Basel, Switzerland.
[5] Corresponding author: helge.grosshans@fmi.ch
[*]: These authors contributed equally.

Abstract

Substrate specificity determines protease functions in physiology and in clinical and biotechnological application. However, affordable and unbiased assays for large-scale quantification of substrate cleavage have been lacking. Here, we develop hiMAPS (high-throughput mapping of protease cleavage sites), a cheap, mass spectrometry-based assay, to derive cleavage motifs for human Dipeptidyl Peptidase Four (DPP4), a key regulator of blood glucose levels. We recapitulate the known benefit of proline in the penultimate (P1) position from the substrate N-terminus, extend it to additional residues, and identify and quantify combinatorial interactions of P1 with its neighbors. These findings reveal extensive cooperativity among the enzyme's active site subsites and allow us to derive a sequence motif that predicts substrate turnover. We show that this information provides new opportunities to engineer stabilized versions of a key substrate of DPP4, the small peptide hormone GLP-1, whose derivatives are used for clinical treatment of type 2 diabetes and obesity. Finally, structural and biochemical characterization of a DPP4 homologue, *C. elegans* DPF-3, reveal specific subsite differences and a distinct cleavage motif, providing insight into the mechanistic basis of the observed specificity. Collectively, our findings present a broadly applicable framework to high-throughput mapping of protease cleavage sites, extend our understanding of protease specificity, and provide a chemical space for substrate engineering of a clinically relevant family of exopeptidases.

Introduction

Despite a reputation for being a promiscuous class of enzymes that indiscriminately digest or degrade proteins, many of the ~600 human proteases exhibit substantial substrate specificity. Thus, they can perform diverse functions in regulation of protein activity, localization, stability and many other molecular and cellular processes important in physiology and disease (Lopez-Otin and Bond 2008). Accordingly, protease inhibitors and substrate agonists have emerged as important classes of drugs not only in antiviral therapies but also for treatment of non-communicable diseases (Florentin, Kostapanos, and Papazafiropoulou 2022; Leung, Abbenante, and Fairlie 2000). For instance, the dipeptidyl peptidase 4 (DPP4) exopeptidase regulates blood glucose metabolism through N-terminal processing of the small peptide hormone GLP-1, which initiates GLP-1 degradation and promotes its renal clearance. DPP4 inhibitors and GLP-1 receptor agonists, i.e., derivatives rendered less sensitive to cleavage by DPP4, have thus become two important classes of drugs for the management of type II diabetes (Deacon 2019; Palmer et al. 2016). However, determination of substrate cleavage motifs and identification of physiologically relevant protease substrates have remained major challenges: In the comprehensive MEROPS peptidase database (Rawlings et al. 2018), a total of only 92,216 cleavages are reported for all >5,200 peptidases together, which limits the broader exploitation of the clinical and biotechnological potential of proteases (Vizovisek et al. 2016).

According to the nomenclature introduced by (Schechter and Berger 1967), proteases cleave their substrates by definition between the P1 and the P1' residues (Fig. 1A). Recognition of P1 by a specific active site pocket, generically termed subsite S1, provides substrate sequence specificity to the enzyme. This is illustrated by the S1 subsite of DPP4, which is highly suited to accommodating the unique features of proline, i.e., a pyrrolidine group and a 'kink' in the backbone, thus making Pro in P1 a major specificity determinant for DPP4 substrates (Thoma et al. 2003). Substrate residues further to the N-terminus (termed P2, P3, etc.) or to the C-terminus (P1', P2', etc.) of P1 may be recognized by additional subsites (S2, S3 and S1', S2', respectively). Although this model implies that the interactions between the individual substrate residues and their corresponding subsites are relatively independent of one another, it is evident that constraints in size, geometry or charge may lead to interactions such that certain sequence combinations are favorable, others unfavorable (Ng, Pike, and Boyd 2009; Qi et al. 2019) (Fig. 1A). However, elucidating such interactions, and thus subsite cooperativity, requires quantitative information on thousands of cleavage events. This has been difficult to obtain even with recent advances in the use of synthetic peptide libraries (O'Donoghue et al. 2012). Greater coverage can be achieved by DNA-encoded peptides using phage display-related or reporter-based assays (e.g., (Chen, Yim, and Bogyo 2019; Koren et al. 2018)) but relies on proxies of cleavage activity such as barcode enrichments or steady-state reporter abundance.

Here, we develop a highly parallelized, mass-spectrometry-based approach to identify substrate motifs for the human DPP4 dipeptiyl peptidase (Walter, Simmons, and Yoshimoto 1980). This exopeptidase of the DPPIV family clips dipeptides off the N-termini of its substrates (Mentlein 1999), with a strong preference for cleavage after proline in the P1 position (Lambeir et al. 2003). Whether and to what extent additional

substrate positions contribute to the specificity for DPP4-mediated cleavage has remained unclear. Structural studies have failed to identify specific constraints for these residues (Rasmussen et al. 2003; Thoma et al. 2003), and functional assays have yielded conflicting data, likely due to small numbers of tested substrates and a lack of systematic analysis (Lambeir et al. 2003).

Assessing cleavage of tens of thousands of distinct peptides obtained from a cheap source, HeLa cell tryptic lysates, in a procedure that we term high-throughput MApping of Protease cleavage Sites (hiMAPS), allows us to identify substrate features that promote or impair cleavage, including combinatorial interactions among different substrate positions. The resulting substrate motifs can predict protein half-lives measured in independent studies, and guide approaches to engineer substrates with altered stability. We validate the specificity of the motifs through applying hiMAPS to an evolutionarily distant hDPP4 homologue, *C. elegans* DPF-3, an orthologue of human DPP8/9, which reveals a related but distinct cleavage motif. Through a DPF-3 structure that we solve using cryogenic electron microscopy (cryo-EM), we can rationalize differences in the subsite interactions between the two exopeptidases. Thus, our results broaden our understanding of DPPIV family proteins by revealing similarities and differences. They provide an unusually comprehensive dissection of subsite cooperativity in proteases and a basis for more targeted approaches to both physiological substrate identification and, as we show in a proof-of-principle experiment for GLP-1, substrate engineering. We propose that hiMAPS can be readily adapted for the characterization of other proteases.

Results

A highly parallelized protease assay identifies beneficial residues in the P1 position of DPP4 substrates

As an important clinical target, hDPP4 has been characterized extensively in research extending over more than three decades. Yet, the peptidase database MEROPS lists only 29 observed cleavages by hDPP4 (https://www.ebi.ac.uk/merops/cgi-bin/pepsum?id=S09.003;type=P, accessed 2 June, 2023). This dataset supports a clear preference for Pro in P1 (21/29 events) and, to a much lower degree, Ala (4/29 events), but it does not reveal further sequence features that could contribute to specificity, and no clear picture has emerged from the assessment of individual substrates (Lambeir et al. 2003). This level of apparent specificity appears surprisingly low, even when considering a possible role of structural features in restricting cleavage to small peptides (Rasmussen et al. 2003). Moreover, many of the known substrates emerged from candidate testing rather than unbiased approaches, leaving it unclear how representative the data is.

To generate an unbiased and much larger dataset, potentially capable of revealing combinatorial motifs of sequence specificity and identifying sequence feature both beneficial and detrimental to cleavage, we developed hiMAPS (Fig. 1B). We assayed the activity of recombinant, commercially available human hDPP4 on a large library of peptides. Specifically, we utilized HeLa cell lysate digested with trypsin as a cheap and diverse source of peptides, from which we could quantify a total of 53,499 unique peptides with a high level of correlation among three replicates (Suppl. Table 1, Suppl. Fig. 1A). Based on the known preference of trypsin for cleavage after arginine or lysine, we designated 43,314 peptides (80.96% of detected) as tryptic based on their matching

any one of the following three sets of criteria; i) peptides start downstream of, and end with, Arg/Lys, ii) peptides start downstream of Arg/Lys and derive from the carboxy terminus of a protein, or iii) peptides end with Arg/Lys and derive from the amino terminus of a protein. These are the input peptides. The remaining 10,185 (19.04%) non-tryptic peptides represent potential products of the digestion with DPP4. Supporting this assignment, a comparison of peptide intensities before and after digestion revealed that most of the tryptic peptide intensities were unaffected or reduced after protease incubation, while most non-tryptic peptides showed substantial increase in signal intensity (Fig. 1C).

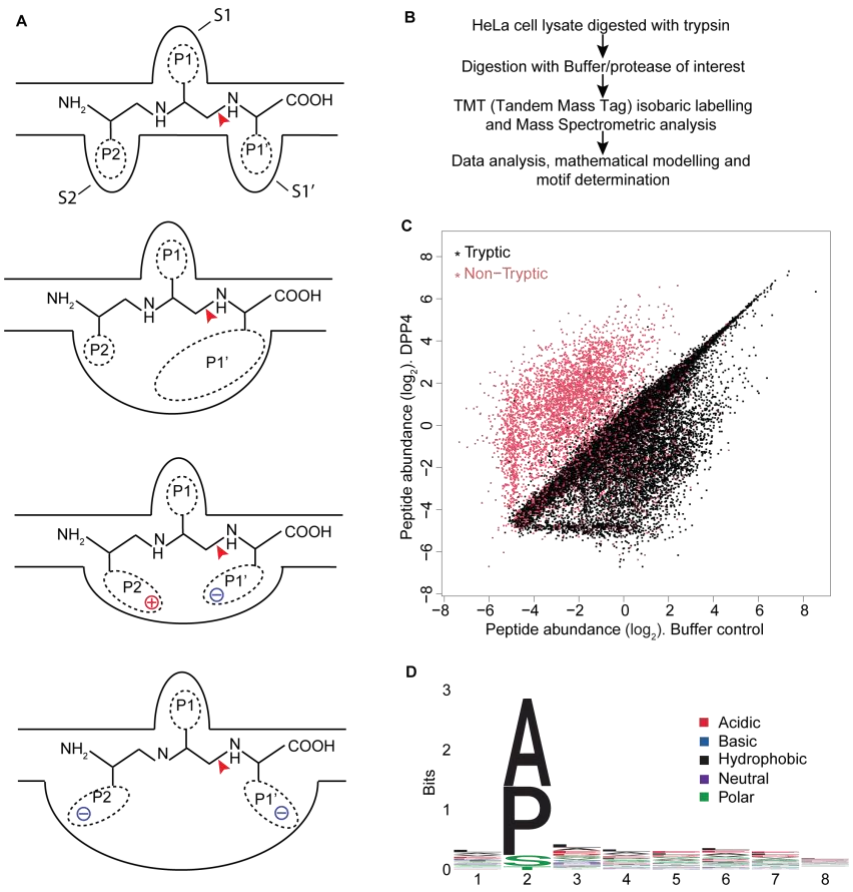


Fig. 1. Determination of DPP4 cleavage motif

A. Cartoon depicting potential interactions among active site sub-sites (S2, S1, S1') and substrate positions (P2, P1, P1'). A red arrowhead indicates the scissile bond.

B. Schematic of hiMAPS.

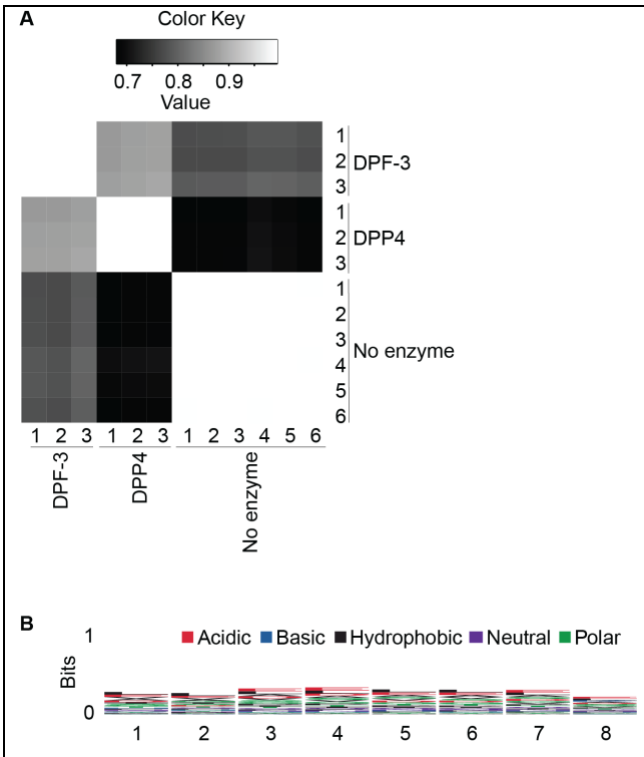
C. Scatter plot showing $\log_2$-transformed peptide abundance in DPP4-treated versus buffer control lysates. Each dot represents a unique peptide. Tryptic and non-tryptic peptides are colored in black and red, respectively.

D. Sequence logo of 1,727 peptides that showed a significant (> 10-fold) change in abundance upon incubation with DPP4.

To base our analyses on faithful and robust quantification of changes for a diverse peptide pool, we focused in the following on the behaviors of substrate rather than product peptides. Thereby, we reduce issues associated with both biases in the detectability of individual peptides from complex mixtures (Mallick et al. 2007) and the lack of baseline levels to which to relate product peptide amounts. We also avoid confounding effects arising from conceptual extension of peptides beyond the true substrate N-terminus, which, in the case of the tryptic lysates that we employ, would yield an artifactual signature of arginine and lysine three amino acids upstream of the cleavage site. Thus, to begin identifying relevant sequence features, we generated a sequence logo aggregating sequence information on the first eight positions of all substrate peptides exhibiting a ≥10-fold change (1,727 peptides). This revealed the identity of the amino acid at the second position (corresponding to P1) as a key feature, with a clear preference for alanine or proline in this position (Fig. 1D).

The fact that we observed similar benefits for alanine and proline in P1 was unexpected given that proline had been reported to outperform any other residue in P1 regarding both substrate fit to the S1 subsite (Thoma et al. 2003) and cleavage (Lambeir et al. 2003). However, it is consistent with the fact that key physiological substrates of DPP4, namely GIP-1 and GLP-1_{7-37}, contain alanine in P1. Alanine contains only a methyl group as a side-chain, which provides two features potentially beneficial for cleavage by DPP4: a small size that is well tolerated in the S1 subsite and a reduction of allowed backbone torsion angles, which facilitates formation of a kinked backbone conformation (Aertgeerts et al. 2004; Rasmussen et al. 2003; Thoma et al. 2003).

Consistent with previous observations (Martin et al. 1993), Ser and Thr also appeared beneficial for cleavage in P1, but to a much lower extent than Ala or Pro. This may be explained by their reaching the size tolerance threshold of the S1 subsite (Rasmussen et al. 2003). We were unable to extract additional information over background from any of the other positions (Suppl. Fig. 1B).



Suppl. Fig. 1. Determination of DPP4 cleavage motif
A. Heatmap showing the correlation of technical replicates.
B. Sequence logo of all tryptic peptides detected in the experiment (background).

hDPP4 target site motifs reveal context-dependency of residues in particular substrate positions

To examine the data in more detail and extract quantitative features, we devised a linear model to predict the tryptic peptide intensity changes in the experiment from their amino acid sequences. To first determine the most important positions within the peptides, we evaluated a series of simple models that used only one particular position (up to the sixth position from the N-terminus) as a single predictor. Consistent with the sequence logo analysis, model fitting showed that the P1 position is by far the most important contributor to enzyme specificity, explaining a substantial 60.7% of the total variance of the peptide changes observed in the experiment (Fig. 2A).

By comparison, the next two most highly scoring positions, P2 and P1', accounted for only 1.4 % and 1.2% of the total variance, respectively. The other three positions (P2'- P4') were essentially negligible at ≤0.5% each, consistent with structural observations of DPP4 in complex with a decapeptide, which revealed that only residues from P2 to P2' form specific interactions with DPP4 side chains (Aertgeerts et al. 2004). Even when we considered the first 3 positions (P2, P1 and P1') simultaneously, in a linear fashion, the predictive power was only slightly increased to 63.1% of total variance. Indeed, the S2 and S1' subsites are relatively open and can accommodate also bulky side chains, explaining the moderate contribution to specificity (Rasmussen et al. 2003; Thoma et al. 2003).

Although the outstanding importance of P1 revealed by our analysis agrees well with the published literature, it left a large fraction of the variance – and thus the features that "make" a substrate – unaccounted. We suspected that certain combinations of residues could be particularly beneficial for, or detrimental to, cleavage. Specifically, a crystal structure of DPP4 in complex with a substrate revealed that the side chains of P2 and P1' point towards each other (Thoma et al. 2003), suggesting that certain rotameric combinations might cause clashing, attraction or repulsion, and thereby affect substrate accommodation in a hydrolysis-competent conformation (Fig. 1A).

The large number of substrates surveyed by hiMAPS allowed us to systematically test the occurrence of this and other interactions. Thus, we created a set of additional models that incorporated interactions terms. Surprisingly, inclusion of P2:P1' interactions increased the predictive power of the model by merely ~0.4 percentage points. By contrast, incorporating interactions between either positions P2 and P1 (P2:P1) or P1 and P1' (P1:P1') increased the predictive power of the model by 3.4 and 6 percentage points, respectively, which was similarly unexpected given that the structural data showed that the P1 side chain is shielded inside the S1 subsite from those of its two neighbors P2 and P1', and that it additionally points in an opposite direction from them (Thoma et al. 2003). Nonetheless, a model including the P2:P1 and the P1:P1' interaction terms simultaneously increased the predictive power of the model by 9.6 percentage points relative to the scenario that considered all positions independently, and by 12 percentage points relative to the "P1 only" baseline scenario. Thus, the final model achieved a predictive power of 72.7%.

The increase in predictive power supports the existence of complex interactions among the three substrate positions that together define substrate specificity. To visualize the rich results of the final model for DPP4, we created multi-panel sequence logos, depicting the contributions of positions P2 and P1' given a certain amino acid in P1 (Fig. 2B, Suppl. Table 2). This allowed us to visualize the full model, including the numerous interaction terms (Methods). The sequence logos depict positive as well as negative contributions of specific amino acids, relative to "average" cleavage, with the value of P1 containing the positional identity term of P1 and the values of P2 and P1' their respective positional identity terms plus their respective interaction terms with P1. Thus, the total predicted cleavage efficiency for a specific amino acid triplet is given by the sum of all the three individual contributions, which may be positive or negative.

Consistent with the analysis above, Pro and Ala in P1 appeared by far the most important cleavage-promoting feature, while Ser and Thr were also beneficial, albeit less so. At the other two positions, Pro in P1' emerged as the most consistent feature, being highly detrimental in the combination with all relevant P1 residues. Indeed, this feature was the largest negative predictor of cleavage from our model, effectively neutralizing even the positive contribution of Pro and Ala at P1. This is explained by the fact that proline at P1' would induce another kink in the peptide backbone which is incompatible with the substrate binding mode. It is also consistent with previous work that failed to observe cleavage of certain P1'(Pro)-containing peptides by hDPP4 (Puschel, Mentlein, and Heymann 1982).

Proline in P1' exhibited a quantitative context-dependent effect in that the size of its negative effect depended on the size of the positive effect of the residue in P1. Other residues in either P2 or P1' revealed even qualitative P1-dependent effects. For instance,

Thr and Val in P1' were mildly beneficial for cleavage when P1 was occupied by Pro but clearly detrimental with Ala, Ser, or Thr in P1. Such context-dependence agrees with a lack of substantial predictive power from individual positions other than P1, or linear combinations of P2, P1, and P1', described above.

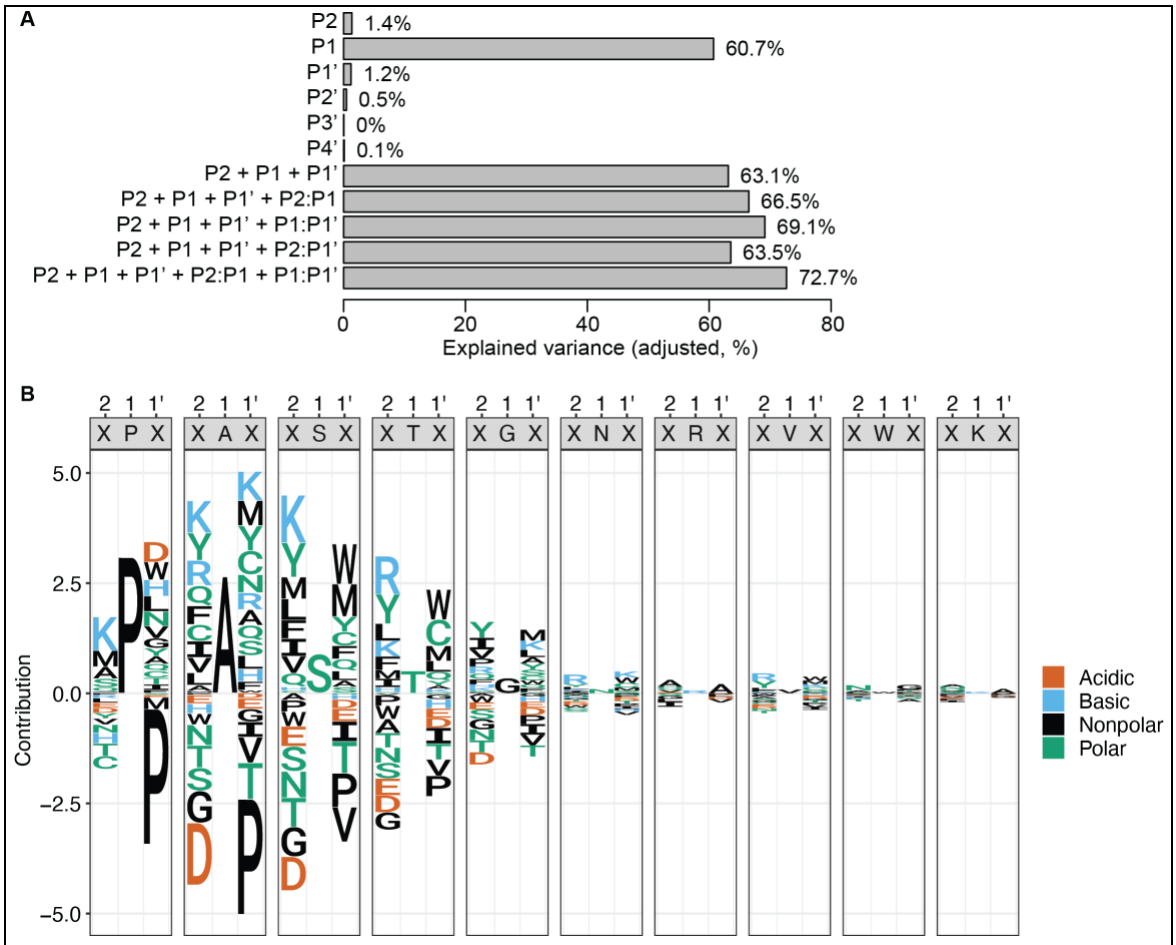


Fig. 2. Models depicting DPP4 cleavage specificity

A. Bar plots showing the predictive power of a total of eleven linear models, each incorporating a specific set of predictors. The included predictors are indicated to the left of the bars. The performance of each model is quantified via the adjusted R^2 to account for differences in the numbers of parameters among the models.

B. Sequence logo representing the cleavage specificity of DPP4 inferred by a linear model. For each amino acid at position P1, a sequence logo triplet depicts the cleavage specificity considering the contributions of P1 as well as the two neighbouring (P2 and P1') amino acids. The height of each amino acid letter in the sequence logo (y-axis) represents the contribution to cleavage efficiency. Positive numbers indicate an increase in cleavage efficiency while negative numbers denote a reduction in cleavage efficiency. The total efficiency of a triplet is given by the sum of the contributions of the amino acids occupying three (P2, P1 and P1') positions.

The DPP4 model predicts substrate turnover kinetics

Given that the subsite interaction data revealed through our analysis could not be predicted from available structural data (Thoma et al. 2003), we sought to validate our

findings and test whether the hiMAPS model could predict cleavage activity on an unrelated dataset. To this end, we used a published dataset providing the half-lives of 21 synthetic peptides (representing physiological and pharmacological substrates of DPP4) upon incubation with DPP4 *in vitro* (Keane et al. 2011). Strikingly, although our model was derived from analysis of peptide turnover at a single time point, and thus did not contain any kinetic information, we observed very good anticorrelation (R=-0.7) between our model prediction score and the previously reported half-lives (Fig. 3A). Thus, peptides predicted to be good substrates according to a high model score had short half-lives, whereas low-scoring peptides had longer half-lives. Notably, this analysis also confirmed that in the context of extended peptides, alanine in P1 can indeed be as beneficial for cleavage as proline, exemplified by GRF-amide, which starts with the tripeptide YAD, as one of the most rapidly processed substrates of DPP4 *in vitro*.

Comprehensive hiMAPS data provide a basis for protein engineering

Human DPP4 is a key regulator of blood glucose level, which it controls chiefly through processing of the incretin Glucagon-Like Peptide (GLP-$1_{(7\text{-}37)}$). The processed GLP-$1_{(9\text{-}37)}$ peptide exhibits reduced affinity to its receptor and becomes a target of rapid renal secretion and subsequent degradation, reducing its half-life to approximately two minutes (Mentlein, Gallwitz, and Schmidt 1993; Muller et al. 2019). Accordingly, DPP4 inhibitors and GLP-$1_{(7\text{-}37)}$ receptor agonists (GLP-1RA) have become important drugs for treatment of type 2 diabetes (Palmer et al. 2016). To achieve clinical utility, GLP-1RAs such as semaglutide required stabilization against DPP4-mediated cleavage, and modifications that include the replacement of Ala in the P1 position with Gly or 2-aminoisobutyric acid helped to increase the half-life to a maximum of about 165 h (Lovshin 2017). Since further GLP-1RA stabilization would simplify medication schedules that require subcutaneous injections, there is a continued interest in further increasing its half-life without compromising its agonist function. However, the protein engineering space is limited by the fact that the amino terminus of GLP-$1_{(7\text{-}37)}$ is deeply embedded in the GLP-1 receptor transmembrane core and required for full receptor activation (Zhang et al. 2017). Hence, the ability to modulate ‘cleavability’ at additional sites may help to meet the competing requirements of stabilization and maintenance of biological activity.

To provide proof-of-principle that hiMAPS can be used to expand the substrate engineering space on a peptide of interest, we assayed the activity of hDPP4 on different 12 residue-long synthetic peptides representing the N-terminus of GLP-$1_{(7\text{-}37)}$ and variants thereof. Specifically, we investigated the effect of altering P2 (from His to Asp) and/or P1’ (from Glu to Pro), two of the most inhibitory amino acids towards cleavage, as predicted by our model (Fig. 2B). Product formation from the canonical substrate (HAE) in an *in vitro* assay was essentially complete within ≤ 30 minutes (Fig. 3B). By contrast, replacement of His at P2 with Asp (DAE) yielded greatly decreased product levels at all time points and continued product accumulation over 120 minutes, consistent with the model prediction of reduced DPP4 activity on this substrate (Fig. 3B). No measurable product generation was observed in buffer-only control reactions (Fig. 3C). Additionally, and again consistent with the model prediction, replacement of glutamate in P1’ with proline (either alone, in the HAP peptide, or together with aspartate in P2, DAP) completely abrogated product formation (Fig. 3B, C, Suppl. Fig. 2).

The introduction of proline in P1' may destabilize the GLP-1 helix fold, which is important for receptor binding, thus negatively affecting activity. However, molecular modeling, using the GLP-1 structure bound to its receptor described in (Zhang et al. 2021), suggested that aspartate in P2 will likely be well tolerated (Fig. 3D). Hence, these data indicate that hiMAPS can be used to rationally design protease substrates with desired properties.

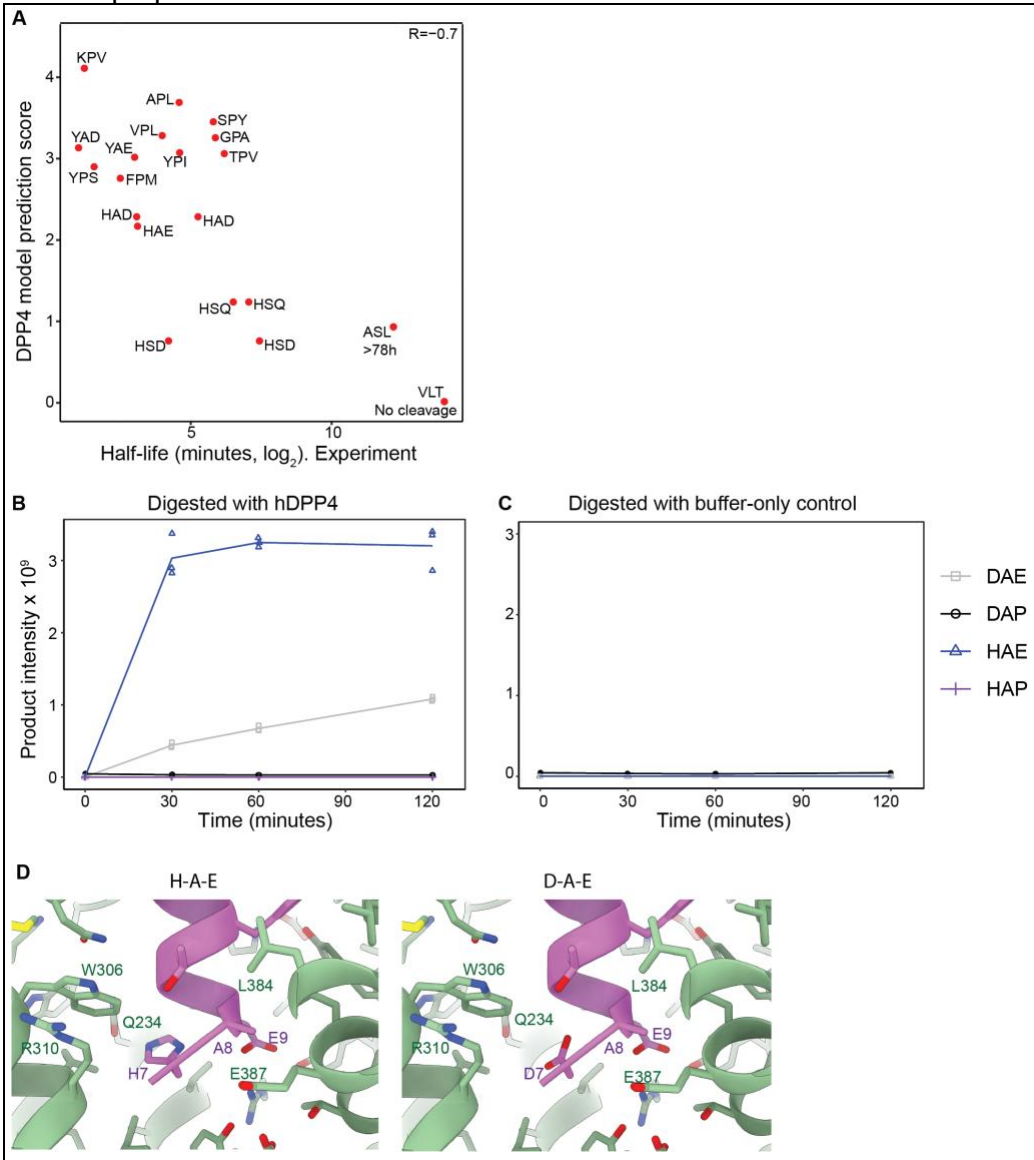
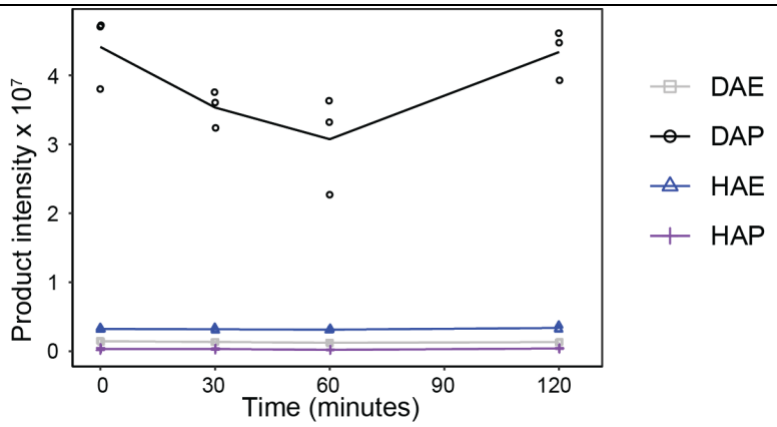


Fig. 3. Models predict substrate turnover rates

A. Scatter plot of experimentally determined peptide half-lives upon incubation with DPP4 (from (Keane et al. 2011)) vs. hiMAPS model substrate score. Each dot is a unique peptide whose first three amino acids are indicated.

B, C. Product peptide intensity (y-axis) of GLP-1 and its variants plotted over reaction time (x-axis) when digested with hDPP4 (B) or buffer-only control (C). The GLP-1 sequence is HAEGTFTSDVSR and is represented as HAE. The variants have changes in the first and/or third residues and are labeled with the first three amino acids. The amino acids 4 through 12 are

common for all the peptides. The data points of three technical replicates are shown; a line indicates mean values.
D. Left panel, N-terminus of GLP-1 (purple) bound to GLP-1 receptor (green, PDB ID 6X18, (Zhang et al. 2020). Right panel, model for GLP-1 peptide with histidine 7 (position P2) replaced by aspartate (structure from left panel used as template, see Methods).



Suppl. Fig. 2. *in vitro* digestion of synthetic peptides
Zoom-in on product peptide intensity (y-axis) of GLP-1 and its variants plotted over reaction time (x-axis) upon incubation with buffer-only control (same experiment as shown in Fig. 3C). Trace levels of the expected product peptides are detected as contaminants of the peptide substrates, confirming that all hypothetical products are detectable in the assay. Note that levels do not change over time and, for HAE and DAE, account for < 1% of signal detected in the presence of enzyme (Fig. 3B).

C. elegans DPF-3 differs in specificity from hDPP4

To examine the substrate specificity of a second DPPIV family exopeptidase, we selected *C. elegans* DPF-3, a DPP8/9-orthologue that promotes male fertility through N-terminal processing of two large RNA-binding proteins, WAGO-1 and WAGO-3 (Gudipati et al. 2021). We repeated hiMAPS using recombinant DPF-3 produced by baculovirus-mediated expression in *Trichoplusia ni* High-Five insect cells. As for DPP4, most non-tryptic peptides showed a substantial increase in signal intensity after incubation with DPF-3 relative to a buffer control, whereas most tryptic peptides had unchanged or decreased levels (Fig. 4A and Suppl. Table 1). A sequence logo analysis covering the most extensively changing substrate peptides (≥10-fold change, 1,133 peptides) again revealed a striking contribution of P1 with a clear preference for Pro and Ala. However, in contrast to DPP4, DPF-3 exhibited substantial preference for Pro over Ala in this position (Fig. 4B). No additional information could be gleaned from any of the remaining positions (Suppl. Fig. 3A).

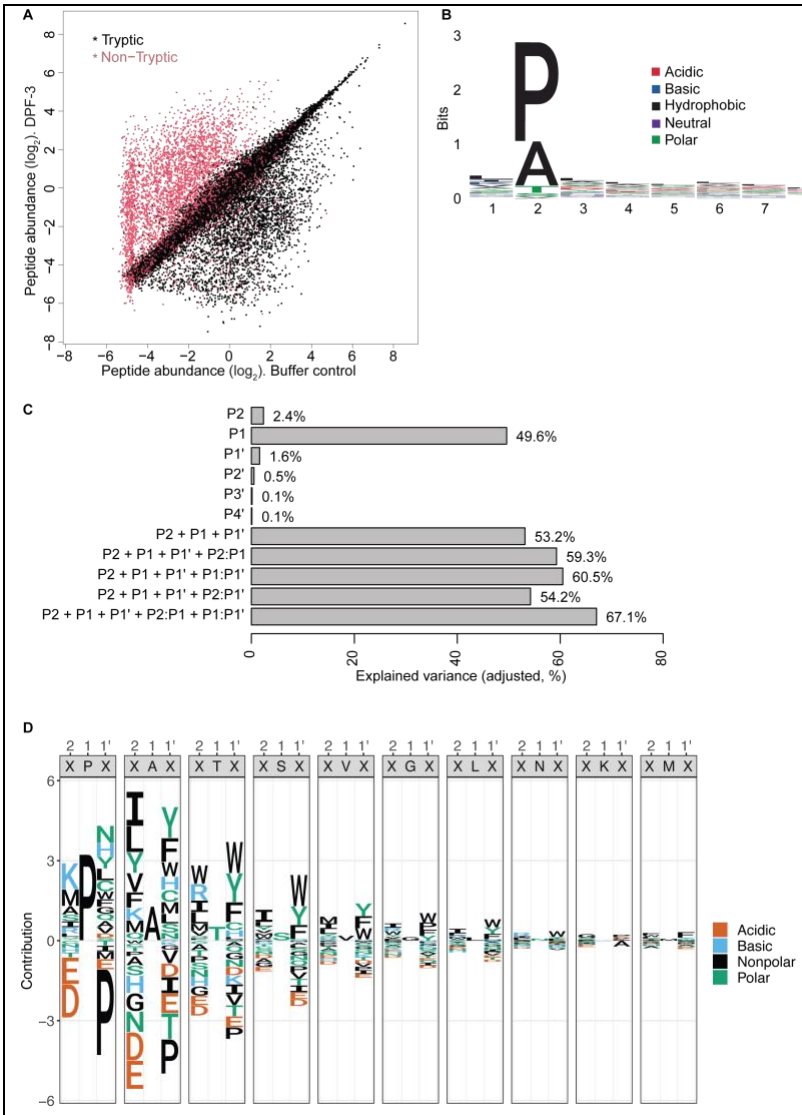


Fig. 4. Determination of the DPF-3 cleavage motif

A. Scatter plot showing $\log_2$-transformed peptide abundance in DPF-3-treated versus buffer control lysates. Each dot represents a unique peptide. Tryptic and non-tryptic peptides are colored in black and red, respectively.

B. Sequence logo of tryptic peptides that showed a significant (> 10 fold) change in abundance upon incubation with DPF-3.

C. Bar plots showing the adjusted predictive power of a total of eleven linear models, each incorporating a specific set of predictors. The included predictors are indicated to the left of the bars. An adjusted R^2 is used to account for differences in parameter numbers among the models.

D. Sequence logo representing the cleavage specificity of DPF-3 inferred by a linear model. For each amino acid at position P1, a sequence logo triplet depicts the cleavage specificity considering the contributions of P1 as well as the two neighbouring (P2 and P1′) amino acids. The height of each amino acid letter in the sequence logo (y-axis) represents the contribution to cleavage efficiency. Positive numbers indicate an increase in cleavage efficiency while negative numbers denote a reduction in cleavage efficiency. For P2 and P1', values comprise both the positional identity and the respective interaction terms with P1; the model intercept term is

integrated into the P1 value. Thus, the total efficiency of a triplet is given by the sum of the contributions of the amino acids occupying the three (P2, P1 and P1′) positions.

We applied the same linear model analysis to DPF-3, thereby revealing a similar set of substrate specificity features, i.e., a large predictive power of P1 (49.6% of total variance), little contribution of any of the other positions individually or jointly, but a substantial increase (to 67.1%) when introducing the two interaction terms P2:P1 and P1:P1' (Fig. 4C, Suppl. Table 3).

In agreement with the results for DPP4, Pro, Ala, Thr and Ser in P1 were all favorable for substrate cleavage (Fig. 4D). However, when we applied the model to predict the half-lives of hDPP4 substrates, it performed substantially less well (R=-0.41 vs. R=-0.7 for the hDPP4 model; Suppl. Fig. 3B). This finding suggested substantial differences between the substrate specificities of DPF-3 and DPP4. Confirming robustness of the results, we repeated DPF-3 hiMAPS (Suppl. Table 4) and obtained highly similar model parameters (R=0.96, Fig. 5A, Suppl. Table 5). By contrast, the correlation was much lower (R=0.76) with the parameters of the DPP4 model.

To visualize the differences in the specificities of DPF-3 and DPP4, we subtracted the model parameters and generated a new sequence logo depicting the differences between the two enzymes. This analysis confirmed the greater preference of DPP4 vs. DPF-3 for Ala, Ser and Gly at P1 (Suppl. Fig. 3C) and revealed extensive differences for various amino acid combinations. Thus, Asp or Glu in P2 were broadly detrimental for DPF-3-mediated cleavage, but much more context-dependent for DPP4 (Fig. 2B, 4D; Suppl. Fig. 3C). For instance, for DPP4, both amino acids appeared largely neutral when Pro occupied P1, yet Asp but not Glu was detrimental when Ala occupied P1. In P1’, both enzymes prefer aromatic or hydrophobic residues, but the effect is generally stronger for DPF-3 than DPP4, and for both enzymes less pronounced when P1 is occupied by Pro than any other amino acid.

Finally, we asked if the two models could detect differences in the specificity of DPF-3 and DPP4 at the single peptide level. To this end, we predicted the cleavage efficiency for all the detected tryptic peptides using both models separately and compared the two output scores (Fig. 5B). This showed that one group of peptides was predicted to undergo efficient cleavage according to both models, while specific subsets of peptides were predicted to undergo preferential cleavage by either DPF-3 or DPP4. Overlaying those predicted subsets of peptides with the experimentally determined abundance changes for each detected tryptic peptide showed a high agreement, indicating that the two models can indeed distinguish the specificity of the two enzymes at the single peptide level.

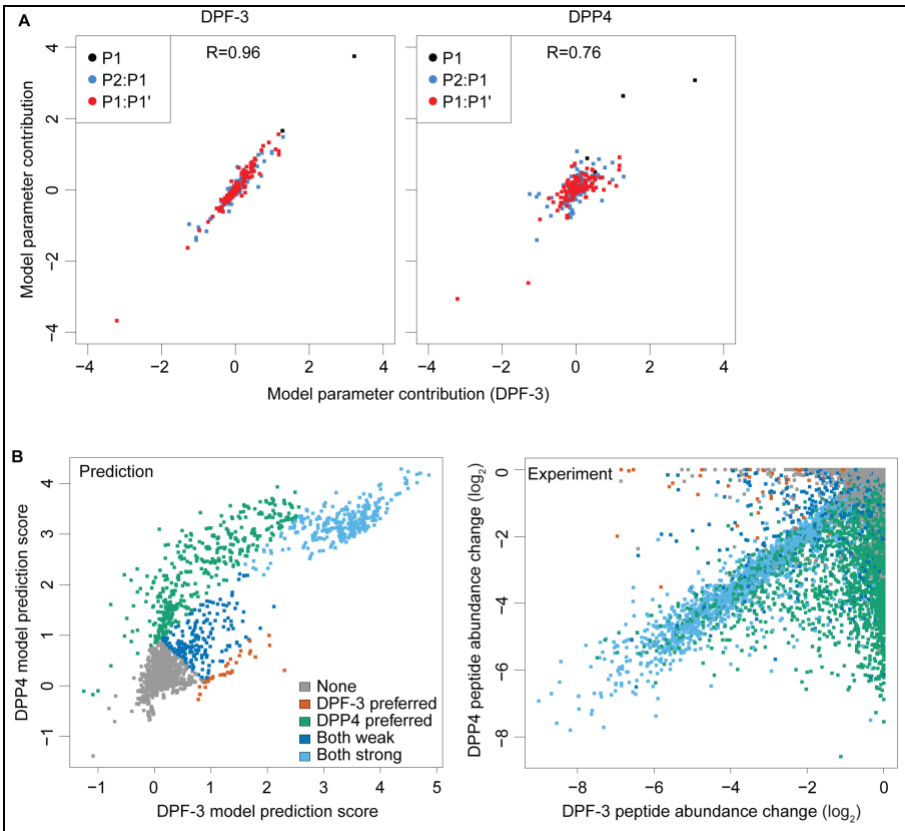
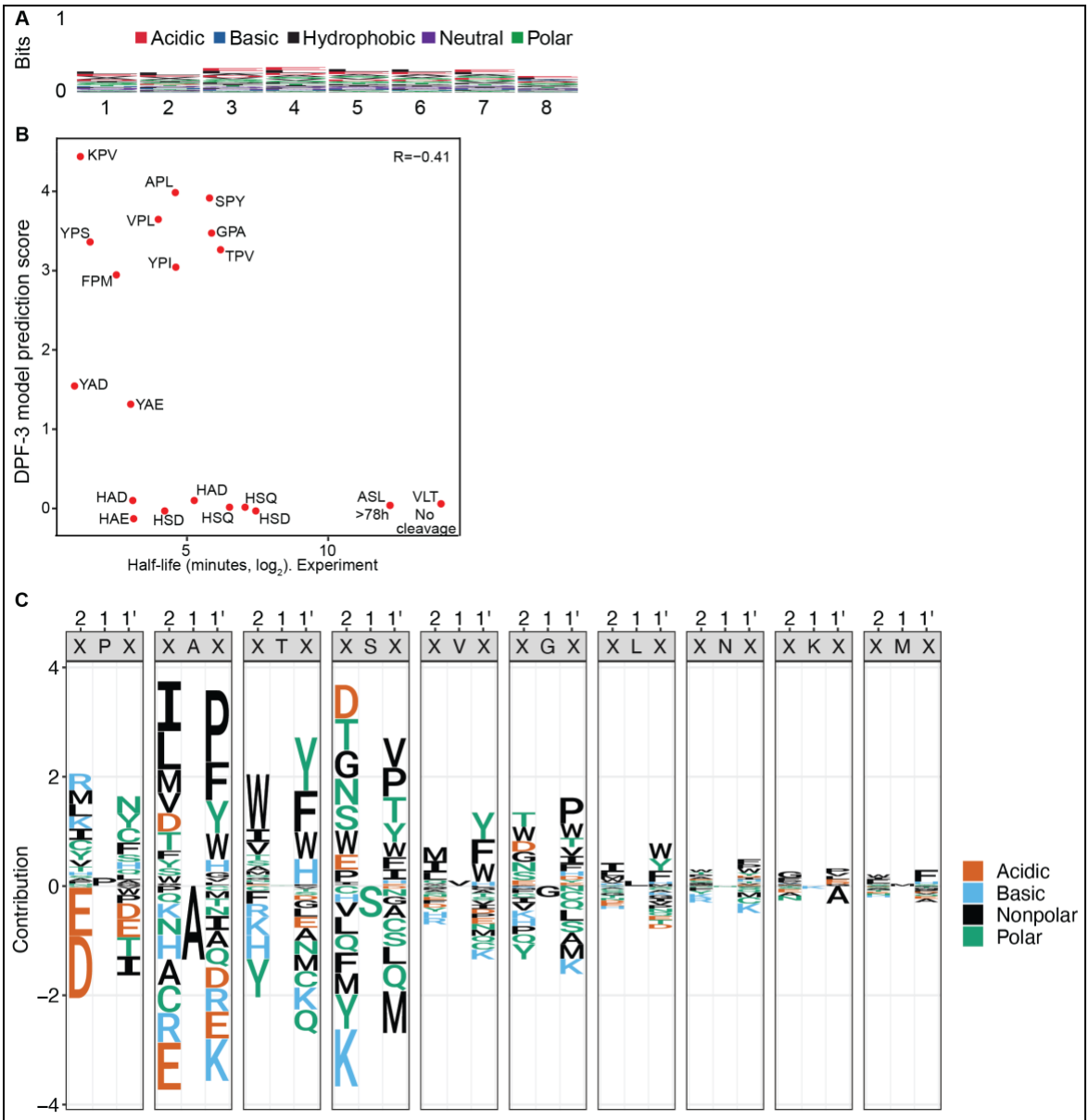


Fig. 5. Substrate preferences of DPF-3 and DPP4 differ

A. Comparison of models generated from an experiment using recombinant DPF-3 (x-axis) and an independent experiment either using DPF-3 (y-axis, left) or DPP4 (y-axis, right). Each dot represents one model parameter.

B. *Left panel*, Scatter plot showing model prediction scores of all experimentally detected tryptic peptides for DPF-3 (x-axis) versus DPP4 (y-axis). The colors denote a visual classification of the peptides as indicated. *Right panel*, Scatter plot showing peptide abundance changes from the experiments. The colors denote the classification derived from the left panel



Suppl. Fig. 3. Determination of the DPF-3 cleavage motif
A. Sequence logo of all tryptic peptides detected in the experiment.
B. Scatter plot of experimentally determined peptide half-lives upon incubation with DPP4 (from (Keane et al. 2011)) vs. DPF-3 hiMAPS model substrate score. Each dot is a unique peptide whose first three amino acids are indicated.
C. Sequence logo showing the differential cleavage activity (DPP4 subtracted from DPF-3) when the P1 position is occupied by a given amino acid in relation to different amino acids occupying P2 and P1' position.

A cryo-EM structure of DPF-3 reveals distinct S2 and S1' subsite features
To relate the observed differences in substrate specificities of DPF-3 and DPP4 to structure, we determined the structure of DPF-3 to an effective resolution of 2.6 Å by single particle cryogenic-Electron Microscopy (cryo-EM). The map was interpretable for

residues 37-123, 259-431, and 470-916. The resolution varied between 2 Å and 6 Å (Fig. 6A and Suppl. Fig. 4, 5). The overall structure follows the conserved dipeptidyl peptidase (DPP) fold that is characterized by two domains, a C-terminal hydrolase with α/β fold, and an N-terminal β-propeller domain encompassing eight blades (Fig. 6B) as observed in other DPPIVs (Ross et al. 2018). DPF-3 superposes with the apo structures of its orthologues DPP8 and DPP9 with r.m.s.d. (Root Mean Square Deviation) values of 3.4 and 3.0 Å (dimer) as well as 3.6 and 2.9 Å (monomer) suggesting strong structural conservation.

A comparison of the DPF-3 apo structure with DPP4 reveals an S1 subsite that is nearly identical for both enzymes, with all amino acids being conserved (Fig. 6C). Moreover, no outstanding differences regarding residue positions or side chain conformations were evident in the structure to explain altered specificities between DPF-3 and DPP4. Instead, the divergent enzymatic footprints appear best explained by differences in the S2 and S1' subsites. Unlike DPP4, DPF-3 contains an acidic residue (E309) but lacks a bulged loop inserting a hydrophobic/aromatic residue (F357 in DPP4) into the S2 subsite, which alters its geometry and chemical environment substantially. E309, which structurally aligns with a serine in DPP4 (S209), is adjacent to the double glutamate motif that is required to position the terminal $NH3^+$ group (Thoma et al. 2003). This difference between the two proteins is consistent with the stronger negative contribution of acidic residues, i.e., Asp and Glu, at P2 in DPF-3 relative to DPP4. A decreased tolerance for certain amino acids in S2 may also require a greater contribution from the other subsites for efficient cleavage, thus offering a potential explanation of DPF-3's clear preference for P1(Pro) with its optimal S1 subsite binding and the constrained substrate backbone.

Another significant difference from DPP4 is the presence of an isoleucine (I705) in the S1' subsite of DPF-3 that is oriented towards the P1' substrate position. In DPP4, the corresponding side chain is a serine (S552), which is retracted from the substrate along with a segment of the loop. This difference is consistent with DPF-3's enhanced preference for several hydrophobic and all aromatic amino acids at P1', because the isoleucine may facilitate hydrophobic packing of these side chains to contribute to substrate binding or positioning.

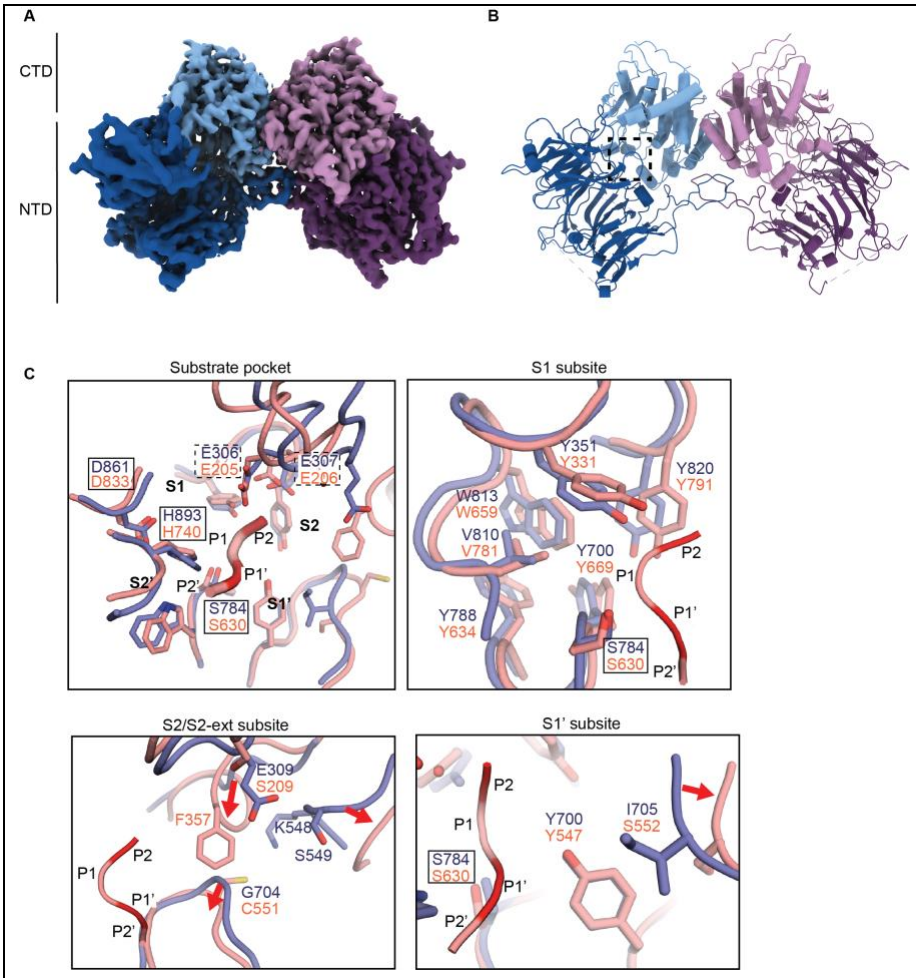
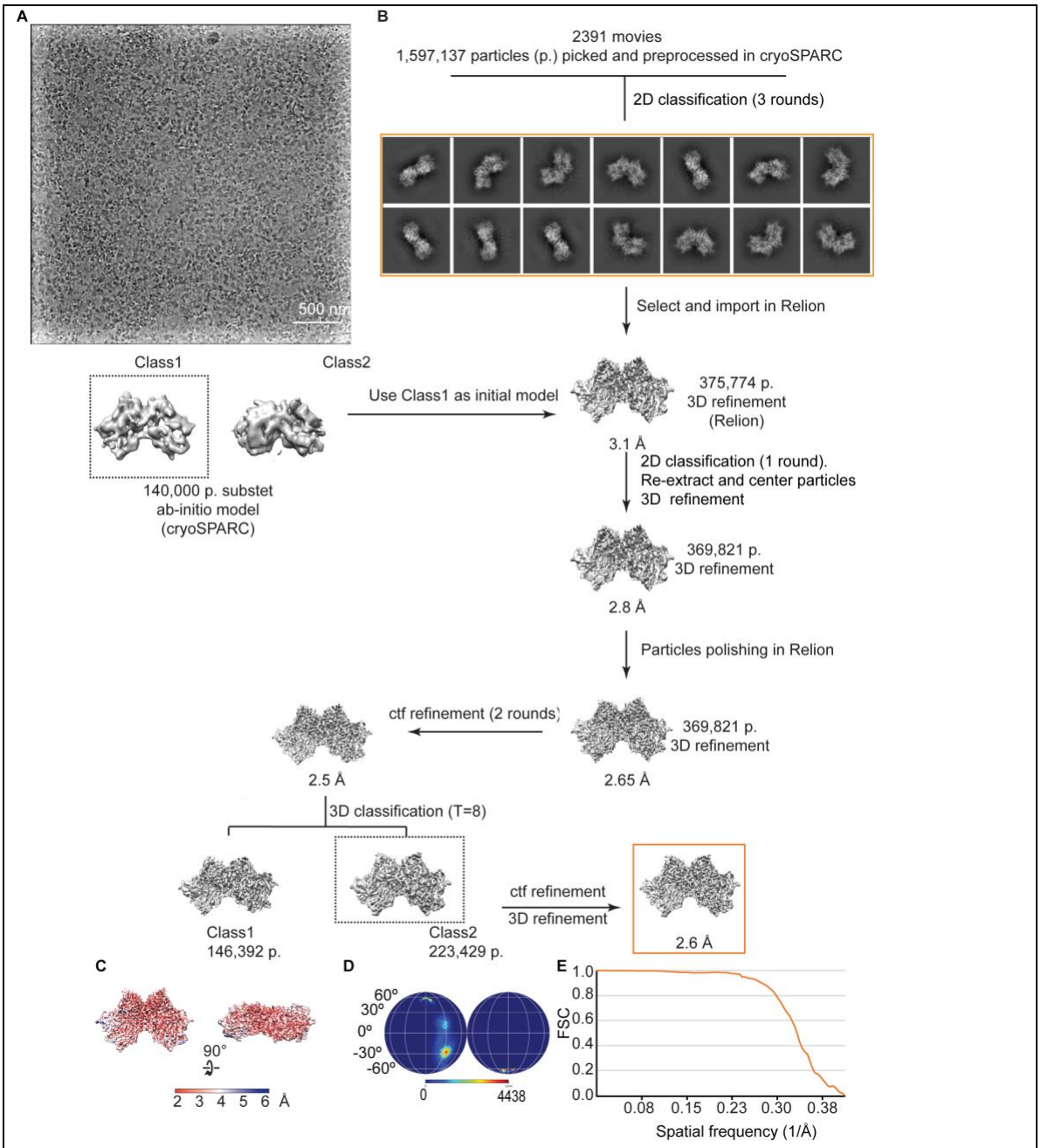


Fig. 6. Cryo-EM structure of DPF-3 and its comparison with hDPP4

A. Cryo-EM structure of dimeric DPF-3 at 2.6 Å. NTD, N-terminal β-propeller Domain (dark shade), CTD, C-terminal α/β-hydrolase domain (light shade).

B. Atomic model derived from the cryo-EM map shown in A. The substrate pocket is boxed.

C. Comparison of DPF-3 substrate pocket with DPP4. DPF-3 (blue) superposed on DPP4 (red) in complex with a substrate (red ribbon, PDB 1R9N). The substrate pocket and the S1, S2, and S1' subsites are shown. Substrate residue positions are indicated with P. Residues of the catalytic triad are boxed. The double glutamate motif is highlighted with dashed boxes. Significant backbone shifts are indicated with red arrows.



Suppl. Fig. 4. Classification and refinement workflow for the DPF-3 homodimer complex

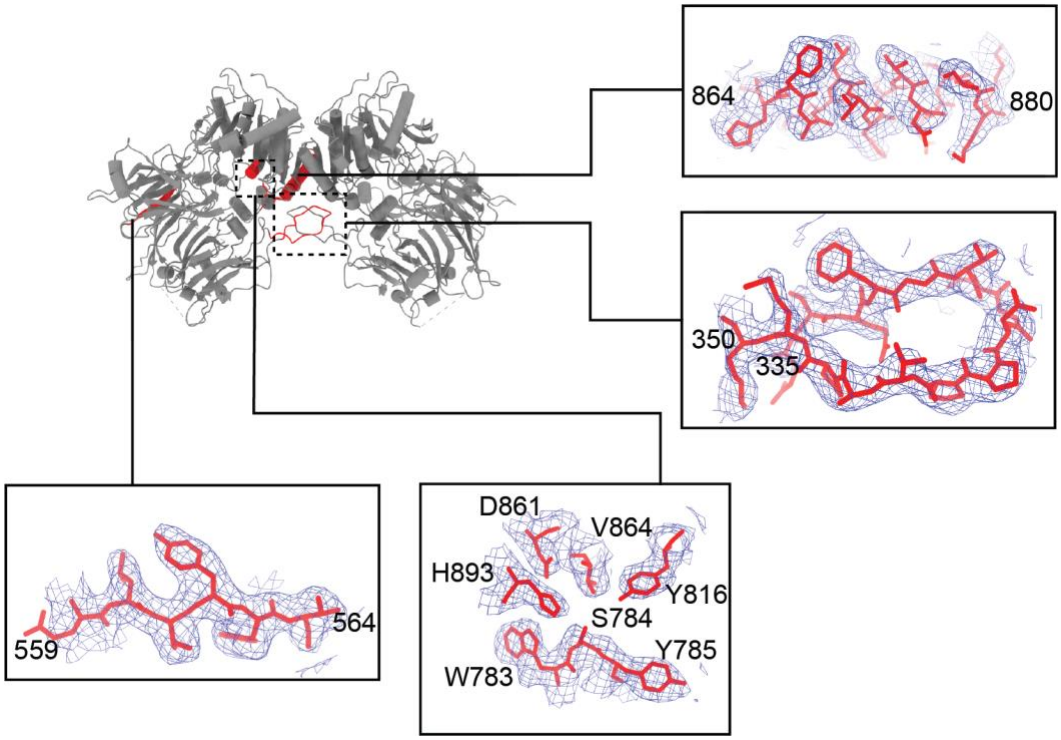
A. Representative cryo-EM micrographs denoised with Topaz, see methods.

B. The movies were imported in cryoSPARC for initial data processing including 2d classification (representative 2d class averages are shown inside an orange frame). The best particles were imported in Relion. Ab initio model generation was performed in cryoSPARC, while 3D classification and refinement were performed in Relion. The final model includes 223,429 particles. The boxes defined by a dashed line indicate the good models and set of particles used for the following step in the data processing workflow.

C. Local resolution filtered map (MonoRes).

D. Angular distribution for the particles leading to the final EM map.

E. Gold-standard Fourier shell correlation curve.



Suppl. Fig. 5. Model map of DPF-3 dimer

Representative model-map overlays (insets) for regions coloured in red in the top left panel. Numbers correspond to sequence positions of residues. The map was sharpened with LocScale and the contour level is 0.192.

Discussion

Knowledge of protease substrates and substrate specificity is a key to understanding both physiological and pathological protease functions, target them for therapeutic use, or optimize the substrates for clinical or biotechnological use through protein engineering approaches. The experimental and analytical procedures that we developed to establish hiMAPS help to generate such knowledge, allowing the identification of combinatorial substrate motifs, including positive and negative contributions, for exopeptidases.

Experimental and analytical advances to identify protease substrate motifs
hiMAPS combines a cheap and readily available source of large numbers of diverse peptides (~$4*10^4$ in our analysis) with direct read-out of cleavage by mass-spectrometry, distinguishing it from other approaches. The combination of deep experimental coverage with mathematical modeling permits hiMAPS to go beyond identification of individually important substrate positions and beneficial residues at that position, to uncovering cooperative effects arising from combinations of residues.

We note that the use of tryptic lysates in the present implementation excludes certain peptides from analysis, e.g., substrates containing arginine or lysine in P1 will be depleted. However, although common mass-spectrometry data analysis tools have been optimized for use of trypsin-digested proteins, use of hiMAPS with other sources of peptides is possible. Thus, lysates can be digested with other proteases such as AspN, GluC, Chymotrypsin, elastase (Dau, Bartolomucci, and Rappsilber 2020) that ensure sufficient diversity, and different sources of biological material can be used for lysate preparation. Indeed, hiMAPS offers users the possibility to select as an input source for lysate preparation the cell type or tissue in which they wish to study the function of their protease of interest, bringing the *in vitro* hiMAPS approach closer to querying *in vivo* protease function.

We expect that hiMAPS can be readily adapted to other exopeptidases, which account for 10 % – 15% of mammalian proteases and whose substrate repertoires as well as cleavage "motifs" frequently remain to be determined (Rawlings et al. 2018). For this subset of proteases, one can exploit the fact that the cleavage site can be anchored to the protein terminus when analyzing substrates. However, we propose that hiMAPS can also be extended to endopeptidases. In this case, identification of matching pairs (or trios) of changing substrates and products would be needed to assign sites of cleavage. Sufficiently large number of quantifiable peptide changes may thus require additional adaptations to sample preparation and data analysis to account for the variable detectability of individual peptides in complex mixtures (Mallick et al. 2007).

Substrate specificity of DPPIV family proteases

Although initially identified through its ability to cleave substrates containing P1(Pro) as reflected in its earlier names glycyl-prolyl-β-naphthylamidase, post-proline dipeptidyl aminopeptidase, and X-Pro dipeptidyl aminopeptidase, it is meanwhile accepted that DPP4 cleaves substrates after residues other than proline (Lambeir et al. 2003). Nonetheless, and given that a preference for P1(Pro) could be well explained by available structural data, we were surprised by the similarity of activity that we observed against substrates containing P1(Pro) and P1(Ala). This is unlikely to be an artifact of hiMAPS since among the five most short-lived substrates whose half-lives were examined by Keane et al., (Keane et al. 2011), two carried P1(Ala), including the least stable substrate, GRF-amide.

Despite a major contribution from P1, its identity does not fully specify activity of DPP4. The extent and way other positions contribute has remained largely unknown, in part because most studies were limited to a relatively small number of substrates, and technological differences prevented comparisons across studies (Lambeir et al. 2003). In our analysis, ~40% of variance in the data remained unaccounted in a model that focused exclusively on P1. Importantly, isolated consideration of additional substrates provided little further information, and only the combination of residues led to a substantial increase in variance explained. In other words, the impact of the identity of P2 or P1' (on which we focused), can only be understood in the context of the identity of P1. We propose that this context dependency prevented the previous, smaller studies from reaching a consensus in understanding the relevance of these residues.

We note that given the architecture of the DPP4 active site, cooperative interactions appear best explained by indirect effects of P2 and P1' on P1. For example,

side chains packing in a specific way against residues in the S2 and S1' subsites could influence the positioning or backbone geometry, thus aiding in or competing with hydrolysis-competent engagement of the scissile bond. Since proline in P1 on the other hand has a strong positioning effect for the P2 residue due to its constrained backbone geometry and its superior accommodation in the S1 subsite, it may be more compatible than any other residue with the negative effects of certain side chains that are not well tolerated in the S2 subsite. Further changes in the S2 and S1' subsites of DPF-3 relative to DPP4 may then explain why combinatorial effects seem to be even more extensive for DPF-3.

Coverage by our data permitted a near complete survey of the 1,200 dipeptide combinations involving P1:P1', P1:P2 and P2:P1' ($3x20^2$), and these combinations were sufficient to explain most of the experimentally observed substrate specificities. Yet, ~30% of variance in the data remained unexplained. Conceivably, some of these effects may be explained by peptide structures that could limit active site accessibility. It also seems likely that more complex interactions among active subsites may exist, which we cannot explore systematically with the current dataset, where we detect for instance 1,577 of the possible 8,000 tripeptide combinations at sufficient coverage (10 detection events per peptide). Future developments in mass-spectrometry and analysis, along with additional adaptations and improvements in sample preparation can be expected to further expand the utility of hiMAPS to more complex combinations.

Comprehensive hiMAPS data provide a basis for protein engineering

Technological advances in protein purification and synthesis have greatly accelerated the development of therapeutic peptide drugs over the past few decades, yet stabilization against proteolytic degradation has remained a bottleneck for successful *in vivo* application (Wang et al. 2022). Chemical modification can achieve efficient peptide stabilization, but often comes with a trade-off of reduced activity on target (Wang et al. 2022). Our proof-of-principle experiments suggest that hiMAPS analysis can extend the peptide engineering space, thereby providing further options for the design of stabilized and biologically active peptides.

We note that such engineering efforts are not necessarily limited to peptide stabilization. For instance, it may be beneficial to render specific peptides sensitive against particular proteases to refine or constrain their spatial and temporal activity patterns or modulate, through processing, the type of activity. In the future, through application of hiMAPS to other proteases, it may be possible to obtain a compendium of relevant cleavage sites that can be used to engineer synthetic substrates de novo, furthering exploitation of protease-mediated processing for synthetic biology.

Experimental details

hiMAPS sample preparation and Mass spectrometry analysis to determine the DPF-3 and DPP4 motifs

The tryptic Pierce HeLa Protein Digest Standard (ThermoFischer Scientific, 88329) was used as a diverse source of tryptic peptides and incubated with the enzyme of interest in HEPES buffer (HEPES pH 7.4 20mM, NaCl 150mM, 2mM TCEP). Assays were

performed in 100 µl volume comprising 10 µg of HeLa protein digest standard and 5 µg of recombinant DPF-3 (see below) or 0.5 µg of DPP4 (Abcam ab79138). The reactions were incubated at 21°C for 4 h followed by heat inactivation at 95°C for 3 minutes. TMT labelling and de-salting was done using the PreOmics iST-NHS kit (Cat # P.O.00026) as described in (Challa et al. 2021). The digested samples were injected, loaded, desalted, and then separated on a 50 cm µPAC C18 HPLC column (Pharmafluidics) connected to a modified Digital PicoView nano-source (New Objective). The following chromatography method was used: 0.1% formic acid (buffer A), 0.1% formic acid in acetonitrile (buffer B), gradient 100 min in total, flow rate 800 nL/min (up to 25 min), then 500 nl/min (25-100 min), mobile phase compositions in % B: 0–5 min 2–7%, 5–25 min 3–6%, 25-30 min 6-8%, 30-70 min 8-20%, 70-88 min 20-32%, 88-89 min 32-95%, 98-96 min 95%, 96-97 min 95-2%, 97-100 min 2%. An Orbitrap Fusion Lumos Tribrid mass spectrometer was operated in a data-dependent mode to quantify TMT reporter ions using synchronous precursor selection-based MS3 fragmentation, as described in (McAlister et al. 2014). Briefly, every 3 seconds, the most intense precursor ions from Orbitrap survey scans (MS1) were selected for collision-induced dissociation fragmentation with a fixed collision energy set to 32%. MS2 CID spectra were generated by the instrument's ion-trap analyzer from which the 10 most abundant notches were selected for MS3 scans. MS3 spectra were recorded using the Orbitrap analyzer at a resolution of 50,000.

MS2 fragment ion spectra were searched with the Sequest HT search engine against the Human Swissprot protein fasta database (downloaded on November 20, 2020). A maximum of two missed cleavages was tolerated for trypsin (semi) enzymatic digestion specificity. Fixed peptide modifications were set for TMTpro / +304.207 Da (Any N-Terminus), TMTpro / +304.207 Da (K), Carbamidomethyl / +57.021 Da (C). Variable peptide modifications were allowed for Oxidation / +15.995 Da (M), Acetyl-noTMTpro / -262.197 Da (N-Terminus). Peptide-to-spectrum matches (PSMs) were validated using the target-decoy search strategy (Elias and Gygi 2007) and Percolator (Kall et al. 2007) with a strict confidence threshold of 0.01, and a relaxed confidence threshold of 0.05. Unique+razor peptides were used for MS3-based TMT6plex reporter-ion quantification, considering Quan Value Correction for isotopic impurities of the TMT pro reagents (Lot# VB294905 and VJ313476), requiring min. PSP mass matches of 65%.

The PeptideGroups table was imported into the peptide workflow in the in-house developed einprot R package version 0.7.0 (Soneson et al. 2023), [https://github.com/fmicompbio/einprot] to undergo $\log_2$ transformation, sample normalization using the center.median approach, and imputation via the MinProb method.

The normalized peptide data was loaded into R as an einprot sce object using the SingleCellExperiment R package (Amezquita et al. 2020), from which peptides with ambiguous N-terminal flanks or with very low detection levels were removed. We used a $\log_2$ intensity threshold of -4.7 for the main experiment (with DPP4 and DPF-3) and a threshold of -4.3 for the DPF-3 only experiment (after inspecting the respective intensity distributions). For peptides with multiple mapping positions, only the first one was considered. To proceed with distinctive peptide sequences for subsequent motif analyses, only the most abundant isoforms of redundant peptides were retained. Data can be downloaded from the Pride with the accession number PXD042089.

hiMAPS data analysis and sequence logo generation
To identify predictive features of cleavage efficiency, we fitted the changes in all tryptic peptides detected by hiMAPS to increasingly complex linear models as detailed in the Results section. Once we had identified and fit the final model that included parameters for the position effects of P2, P1, and P1' as well as interaction effects of P2:P1 and P1:P1', we extracted all the coefficients and rearranged them into derived terms that allowed for simpler interpretability and visualization (i.e., in sequence logos). To do so, we first added back the left-out amino acids at all positions and interaction terms at a value of zero and mean normalized all the coefficients. Secondly, we moved the contributions of P2 and P1' into the interaction terms of P2:P1 and P1:P1' which reduced the total number of terms that need to be added up for every prediction from 6 (Intercept,P2,P1,P1',P2:P1,P1:P1') to only 4 (Intercept,P1,P2:P1,P1:P1'). Thirdly we moved the intercept term and all the shifts caused by the various mean normalization steps into the coefficients of P1. This resulted in a total of 3 terms that need to be added for a given prediction (P1,P2:P1,P1:P1'). Note that two substrates examined by Keane et al., (Keane et al. 2011) were annotated to undergo multiple, successive cleavages; hence, these were excluded from our analysis that examined the correlation between model scores and half-lives since we could not assign a unique experimental half-life to a given tripeptide sequence combination.

Modelling of GLP-1 variant peptide with its receptor
Modelling of a GLP-1 peptide with histidine replaced by aspartate at position 7 was carried out in ChimeraX using the '*Rotamers*' tool. As template, PDB ID 6X18 (Zhang et al. 2021) was used. Clashes for all possible aspartate rotamers (assuming a fixed backbone) were calculated with standard overlap settings (VDW overlap >= 0.6 Å after subtracting 0.4 Å for H-bonding). Two rotamers similar to the histidine conformation showed no clashes, of which one was selected for a representative figure (Fig. 3B).

Production of recombinant DPF-3
Wild-type or S784A mutated DPF-3 cDNA was generated from RNA extracted from N2 (wt) and *dpf-3(xe71[S784A])* mutant animals, respectively. The cDNA was used to amplify and clone the DPF-3 into the pAC8 vector using Gibson assembly. DNA oligonucleotides used in this study are listed in Suppl. Table 8. Strep tag was inserted just before the translation stop codon of DPF-3. Truncated (amino acids 216-252 are replaced by 2x Gly Ser) DPF-3 constructs were produced using Q5 Hot Start High-Fidelity 2x Master Mix (New England Biolabs, NEB # E0554S) followed by the KLD reaction according to manufacturer's instructions. The proteins were expressed and purified exactly as described (Kassube and Thoma 2020) except that the *Trichoplusia ni* High-Five insect cells expressing the desired proteins were grown at 16°C instead of 27°C. Briefly, *Spodoptera frugiperda* (Sf9) insect cells grown in EX-CELL 420 medium (Sigma) were used to produce Baculoviruses through recombination by cotransfection of the pAC-derived vector having the gene of interest and viral DNA using Cellfectin II Reagent (Catalog number: 10362100, ThermoFisher Scientific). The viruses were amplified till Passage 3 (P3) in Sf9 insect cells and the *Trichoplusia ni* High-Five insect cells were used for large scale protein expression and purification. Post infection, the cells were allowed to grow at 16 °C for 5 days at 120 r.p.m, collected and lysed by

sonication in buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 1 mM TCEP, 2 mM PMSF and cOmplete EDTA-free protease inhibitor tablets (Roche, 1 tablet per 50 ml). The lysate was clarified by ultracentrifugation for 30 min at 40,000*g*. Miracloth (Merck) was used to filter the supernatant before loading onto the affinity column.

In vitro protease assay by mass spectrometry

Synthetic peptides representing either wild type or mutated GLP-1 were purchased from Thermo Fischer Scientific and their cleavage by DPP4 was measured over time. Briefly, the lyophilized peptides were dissolved in TBS buffer supplemented with 2 mM TCEP to obtain a stock solution with a final concentration of 0.1 nmol/µl. 1ug of DPP4 and each peptide at a final concentration of 1 pmol/µl were used in reactions at a final volume of 100 µl. The assay was performed at 21°C, aliquots (20 µl) were taken out at the indicated times and mixed with 180 µl of 0.1 % trifluoroacetic acid, 2 mM TCEP, 2 % acetonitrile in water to stop the reaction.

The peptides were analyzed by capillary liquid chromatography tandem mass spectrometry (LC-MS). Peptides were loaded at 45°C onto a 75 µm×15 cm EasyC18 column at a constant pressure of 800 bar, using an VanquishNeo-nLC 1000 liquid chromatograph with one-column set up (Thermo Scientific). Peptides were separated with a gradient of 0-0.2min 2-6%B in A, 0.2-4.2min 6-21%, 4.2-6.2min 21-30%, 6.2-6.7min 30-36%, 6.7-7.2min 36-45%, 7.2-7.5min 45%-100%, 7.5-12min 100%. A: 0.1%FA in H_2O; B: 0.1%FA, 80% MeCN in H_2O, at room temperature and the flow rate during the gradient was 200nl/min and for the last 4.4min 350nl/min. The column was mounted on a DPV ion source (New Objective) connected to an Orbitrap Fusion LUMOS (Thermo Scientific), data were acquired using 120,000 resolution, for the peptide measurements in the Orbitrap and a top T (1.5 s) method with HCD fragmentation for each precursor and fragment measurement in the ion trap following the manufacturer guidelines (Thermo Scientific). MS1 signals were quantified using Skyline 4.1 (MacLean et al. 2010) to generate the results shown in Fig. 3B,C and Suppl. Fig 2.

Cryo-EM sample preparation

3.5 µl of full length DPF-3 sample (~0.3 mg/ml and 2.5% glycerol) was applied to Quantifoil holey carbon grids (R 1.2/1.3 200-mesh, Quantifoil Micro Tools). Glow discharging was carried out in a Solarus plasma cleaner (Gatan) for 15 sec in a H_2/O_2 environment. Grids were blotted for 3-4 s at 4°C at 100% humidity in a Vitrobot Mark IV (FEI, Hillsboro, OR, USA), and then immediately plunged into liquid ethane.

Cryo-EM data collection

Data were collected automatically with EPU (Thermo Fisher Scientific) on a Cs-corrected (CEOS GmbH, Heidelberg, Germany) Titan Krios (Thermo Fisher Scientific) electron microscope at 300 keV. A total of 2391 movies were recorded using the Falcon4 direct electron detector (Thermo Fisher Scientific). The acquisition was performed at a magnification of 75,000x yielding a pixel size of 0.845 Å at the specimen level. The dose rate was set to 9.5 e^-/Å/s and the exposure time was adjusted to account for an accumulated dose of 50 $e^-/Å^2$. The EER frames were grouped to obtain 50 fractions. The targeted defocus values ranged from -0.8 to -2.5 µm.

Cryo-EM image processing

Real-time evaluation along with acquisition with EPU (Thermo Fisher Scientific) was performed with *CryoFLARE* (Schenk et al. 2020). Drift correction was performed with the *Relion 3* motioncorr implementation (Zivanov et al. 2018), where a motion corrected sum of all frames was generated with and without applying a dose weighting scheme. The CTF was fitted using *GCTF* (Zhang 2016). All the motion corrected micrographs that showed a CTF estimation better than 4 Å resolution were imported into cryoSPARCv3 (Punjani et al. 2017) for further processing. After patch CTF and blob picker particle picking, 3 rounds of 2d classification were used to obtain a subset of 375,774 particles that were imported into Relion for further processing. A random subset including 140,000 particles were used to generate two *ab-initio* models in cryoSPARCv3, one of each was used as initial model in Relion.

A combination of 3D classification, 3D refinement, particle polishing and CTF refinement in Relion (Suppl. Fig. 4) led to a DPF-3 map at 2.6 Å resolution. The resolution values reported for all reconstructions are based on the gold-standard Fourier shell correlation curve (FSC) at 0.143 criterion (Rosenthal and Henderson 2003; Scheres 2012) and all the related FSC curves are corrected for the effects of soft masks using high-resolution noise substitution (Chen et al. 2013). All local resolutions were estimated with *MonoRes* (XMIPP) (de la Rosa-Trevin et al. 2013).

Cryo-EM model building, refinement and validation

As initial model, a homology model for DPF-3 was obtained from Swissprot and docked into the structure using CHIMERAX (Pettersen et al. 2021) 'fit-in-map' tool followed by flexible-fitting using ISOLDE (Croll 2018). The model was further refined by iterative rounds of manual rebuilding in COOT (Emsley et al. 2010) and ISOLDE, followed by minimization using the ROSETTA FastRelax protocol in combination with density scoring (Wang et al. 2016). For the latter step, an in-house pipeline was used to automate the ROSETTA protocol (Georg Kempf, 2021: https://github.com/fmi-basel/RosEM). Refinement in ROSETTA was done against the first half-map and the map-model FSC was compared to the map-model FSC of the second half map to test for overfitting. During the course of this study, ALPHAFOLD 2.0 (Jumper et al. 2021) became available and the predicted structure was used for cross-validation of medium-resolution segments. At the final stage, B-factors were fitted using ROSETTA. The structure was validated with MOLBROBITY (Williams et al. 2018), PHENIX (Afonine et al. 2018) and EMRINGER (Barad et al. 2015). Local amplitude-scaling was performed with LocScale (Jakobi, Wilmanns, and Sachse 2017).

Acknowledgements

We thank Iskra Katic, Anca Neagu and Lan Xu for technical support in the generation of transgenic animals. We would like to thank Ganesh Pathare for his kind help and suggestions with regard to the production of recombinant protein and Alicia Michael, Joscha Weiss and Georg Petzold for helpful discussions. We thank Kathrin Braun and Thomas Welte for their inputs and stimulating discussions. We thank Nicolas Thomä and Guillaume Diss for a critical reading of the manuscript, and Nicolas Thomä for access to laboratory resources. We are grateful to Vytautas Iesmantavicius for help with peptide fractionation and recording of LC-MS data. Some strains were provided by the *Caenorhabditis* Genetics Center (CGC), which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440).

Funding
This work was supported as a part of the NCCR RNA & Disease, a National Centre of Excellence in Research, funded by the Swiss National Science Foundation (grant number 182880), by the Novartis Research Foundation through the Friedrich Miescher Institute for Biomedical Research (to H.G.) and by the National Science Center (NCN), Poland, SONATA BIS 2021/42/E/NZ1/00336 and OPUS 2022/45/B/NZ2/02183 (to R.K.G).

Author Contributions
RKG and HG conceived the project, designed and analyzed the experiments and wrote the manuscript with critical inputs from the rest of the authors. HG supervised the project. The experiments related to solving the structure of DPF-3 were performed by SM, SC, and GK. SC and GK drafted the sections related to the structure. JS performed mass spectrometry for hiMAPS. DH did the mass spectrometry for protease assays on specific synthetic peptides. DG conceived and performed the analysis of the mass spectrometry data related to DPF-3 and DPP4 hiMAPS.

Competing interests

The authors declare no competing interests

Data and material availability

The electron cryo-microscopy map and model coordinates were deposited in the Electron Microscopy Data Bank (EMDB) (accession code: EMD-17582) and in the Protein Data Bank (PDB) (accession code: 8PBA). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al. 2022) partner repository with the dataset identifier PXD042089. Published research reagents from the FMI are shared with the academic community under a Material Transfer Agreement (MTA) having terms and conditions corresponding to those of the UBMTA (Uniform Biological Material Transfer Agreement). The einprot package is available from GitHub (https://github.com/fmicompbio/einprot). Archives of the versions used here are available from https://doi.org/10.5281/zenodo.8031403 (v0.6.8) and https://doi.org/10.5281/zenodo.8031408 (v0.7.0).

Supplementary Data
Suppl. Tables 1-5 contain the data related to hiMAPS. Suppl. Table 6 provides a list of the primers used in this study. Suppl. Table 7 provides the cryo-EM data.

Suppl. Table 1: Peptides detected from Tryptic HeLa cell lysate digested with buffer/DPF-3/DPP4.
Suppl. Table 2: Model parameters of DPP4.
Suppl. Table 3: Model parameters of DPF-3.
Suppl. Table 4: Peptides detected from Tryptic HeLa cell lysate digested with buffer/DPF-3 (replicate).
Suppl. Table 5: Model parameters of DPF-3 (replicate).
Suppl. Table 6: Primers used in this study.

Suppl. Table 7: Collection and refinement of apo-DPF-3 cryo-EM data.

References

Aertgeerts, K., S. Ye, M. G. Tennant, M. L. Kraus, J. Rogers, B. C. Sang, R. J. Skene, D. R. Webb, and G. S. Prasad. 2004. 'Crystal structure of human dipeptidyl peptidase IV in complex with a decapeptide reveals details on substrate specificity and tetrahedral intermediate formation', *Protein Sci*, 13: 412-21.

Afonine, P. V., B. P. Klaholz, N. W. Moriarty, B. K. Poon, O. V. Sobolev, T. C. Terwilliger, P. D. Adams, and A. Urzhumtsev. 2018. 'New tools for the analysis and validation of cryo-EM maps and atomic models', *Acta Crystallogr D Struct Biol*, 74: 814-40.

Amezquita, R. A., A. T. L. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Soneson, L. Waldron, H. Pages, M. L. Smith, W. Huber, M. Morgan, R. Gottardo, and S. C. Hicks. 2020. 'Orchestrating single-cell analysis with Bioconductor', *Nat Methods*, 17: 137-45.

Barad, B. A., N. Echols, R. Y. Wang, Y. Cheng, F. DiMaio, P. D. Adams, and J. S. Fraser. 2015. 'EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy', *Nat Methods*, 12: 943-6.

Challa, K., C. D. Schmid, S. Kitagawa, A. Cheblal, V. Iesmantavicius, A. Seeber, A. Amitai, J. Seebacher, M. H. Hauer, K. Shimada, and S. M. Gasser. 2021. 'Damage-induced chromatome dynamics link Ubiquitin ligase and proteasome recruitment to histone loss and efficient DNA repair', *Mol Cell*, 81: 811-29 e6.

Chen, S., G. McMullan, A. R. Faruqi, G. N. Murshudov, J. M. Short, S. H. Scheres, and R. Henderson. 2013. 'High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy', *Ultramicroscopy*, 135: 24-35.

Chen, S., J. J. Yim, and M. Bogyo. 2019. 'Synthetic and biological approaches to map substrate specificities of proteases', *Biol Chem*, 401: 165-82.

Croll, T. I. 2018. 'ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps', *Acta Crystallogr D Struct Biol*, 74: 519-30.

Dau, T., G. Bartolomucci, and J. Rappsilber. 2020. 'Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin', *Anal Chem*, 92: 9523-27.

de la Rosa-Trevin, J. M., J. Oton, R. Marabini, A. Zaldivar, J. Vargas, J. M. Carazo, and C. O. Sorzano. 2013. 'Xmipp 3.0: an improved software suite for image processing in electron microscopy', *J Struct Biol*, 184: 321-8.

Deacon, C. F. 2019. 'Corrigendum: Physiology and Pharmacology of DPP-4 in Glucose Homeostasis and the Treatment of Type 2 Diabetes', *Front Endocrinol (Lausanne)*, 10: 275.

Elias, J. E., and S. P. Gygi. 2007. 'Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry', *Nat Methods*, 4: 207-14.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. 'Features and development of Coot', *Acta Crystallogr D Biol Crystallogr*, 66: 486-501.

Florentin, M., M. S. Kostapanos, and A. K. Papazafiropoulou. 2022. 'Role of dipeptidyl peptidase 4 inhibitors in the new era of antidiabetic treatment', *World J Diabetes*, 13: 85-96.

Gudipati, R. K., K. Braun, F. Gypas, D. Hess, J. Schreier, S. H. Carl, R. F. Ketting, and H. Grosshans. 2021. 'Protease-mediated processing of Argonaute proteins controls small RNA association', *Mol Cell*, 81: 2388-402 e8.

Jakobi, A. J., M. Wilmanns, and C. Sachse. 2017. 'Model-based local density sharpening of cryo-EM maps', *Elife*, 6.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.

Kall, L., J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. 2007. 'Semi-supervised learning for peptide identification from shotgun proteomics datasets', *Nat Methods*, 4: 923-5.

Kassube, S. A., and N. H. Thoma. 2020. 'Structural insights into Fe-S protein biogenesis by the CIA targeting complex', *Nat Struct Mol Biol*, 27: 735-42.

Keane, F. M., N. A. Nadvi, T. W. Yao, and M. D. Gorrell. 2011. 'Neuropeptide Y, B-type natriuretic peptide, substance P and peptide YY are novel substrates of fibroblast activation protein-alpha', *FEBS J*, 278: 1316-32.

Koren, I., R. T. Timms, T. Kula, Q. Xu, M. Z. Li, and S. J. Elledge. 2018. 'The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons', *Cell*, 173: 1622-35 e14.

Lambeir, A. M., C. Durinx, S. Scharpe, and I. De Meester. 2003. 'Dipeptidyl-peptidase IV from bench to bedside: an update on structural properties, functions, and clinical aspects of the enzyme DPP IV', *Crit Rev Clin Lab Sci*, 40: 209-94.

Leung, D., G. Abbenante, and D. P. Fairlie. 2000. 'Protease inhibitors: current status and future prospects', *J Med Chem*, 43: 305-41.

Lopez-Otin, C., and J. S. Bond. 2008. 'Proteases: multifunctional enzymes in life and disease', *J Biol Chem*, 283: 30433-7.

Lovshin, J. A. 2017. 'Glucagon-like Peptide-1 Receptor Agonists: A Class Update for Treating Type 2 Diabetes', *Can J Diabetes*, 41: 524-35.

MacLean, B., D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss. 2010. 'Skyline: an open source document editor for creating and analyzing targeted proteomics experiments', *Bioinformatics*, 26: 966-8.

Mallick, P., M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. 2007. 'Computational prediction of proteotypic peptides for quantitative proteomics', *Nat Biotechnol*, 25: 125-31.

Martin, R. A., D. L. Cleary, D. M. Guido, H. A. Zurcher-Neely, and T. M. Kubiak. 1993. 'Dipeptidyl peptidase IV (DPP-IV) from pig kidney cleaves analogs of bovine growth hormone-releasing factor (bGRF) modified at position 2 with Ser, Thr or

Val. Extended DPP-IV substrate specificity?', *Biochim Biophys Acta*, 1164: 252-60.

McAlister, G. C., D. P. Nusinow, M. P. Jedrychowski, M. Wuhr, E. L. Huttlin, B. K. Erickson, R. Rad, W. Haas, and S. P. Gygi. 2014. 'MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes', *Anal Chem*, 86: 7150-8.

Mentlein, R. 1999. 'Dipeptidyl-peptidase IV (CD26)--role in the inactivation of regulatory peptides', *Regul Pept*, 85: 9-24.

Mentlein, R., B. Gallwitz, and W. E. Schmidt. 1993. 'Dipeptidyl-peptidase IV hydrolyses gastric inhibitory polypeptide, glucagon-like peptide-1(7-36)amide, peptide histidine methionine and is responsible for their degradation in human serum', *Eur J Biochem*, 214: 829-35.

Muller, T. D., B. Finan, S. R. Bloom, D. D'Alessio, D. J. Drucker, P. R. Flatt, A. Fritsche, F. Gribble, H. J. Grill, J. F. Habener, J. J. Holst, W. Langhans, J. J. Meier, M. A. Nauck, D. Perez-Tilve, A. Pocai, F. Reimann, D. A. Sandoval, T. W. Schwartz, R. J. Seeley, K. Stemmer, M. Tang-Christensen, S. C. Woods, R. D. DiMarchi, and M. H. Tschop. 2019. 'Glucagon-like peptide 1 (GLP-1)', *Mol Metab*, 30: 72-130.

Ng, N. M., R. N. Pike, and S. E. Boyd. 2009. 'Subsite cooperativity in protease specificity', *Biol Chem*, 390: 401-7.

O'Donoghue, A. J., A. A. Eroy-Reveles, G. M. Knudsen, J. Ingram, M. Zhou, J. B. Statnekov, A. L. Greninger, D. R. Hostetter, G. Qu, D. A. Maltby, M. O. Anderson, J. L. Derisi, J. H. McKerrow, A. L. Burlingame, and C. S. Craik. 2012. 'Global identification of peptidase specificity by multiplex substrate profiling', *Nat Methods*, 9: 1095-100.

Palmer, S. C., D. Mavridis, A. Nicolucci, D. W. Johnson, M. Tonelli, J. C. Craig, J. Maggo, V. Gray, G. De Berardis, M. Ruospo, P. Natale, V. Saglimbene, S. V. Badve, Y. Cho, A. C. Nadeau-Fredette, M. Burke, L. Faruque, A. Lloyd, N. Ahmad, Y. Liu, S. Tiv, N. Wiebe, and G. F. Strippoli. 2016. 'Comparison of Clinical Outcomes and Adverse Events Associated With Glucose-Lowering Drugs in Patients With Type 2 Diabetes: A Meta-analysis', *JAMA*, 316: 313-24.

Perez-Riverol, Y., J. Bai, C. Bandla, D. Garcia-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, and J. A. Vizcaino. 2022. 'The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences', *Nucleic Acids Res*, 50: D543-D52.

Pettersen, E. F., T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin. 2021. 'UCSF ChimeraX: Structure visualization for researchers, educators, and developers', *Protein Sci*, 30: 70-82.

Punjani, A., J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker. 2017. 'cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination', *Nat Methods*, 14: 290-96.

Puschel, G., R. Mentlein, and E. Heymann. 1982. 'Isolation and characterization of dipeptidyl peptidase IV from human placenta', *Eur J Biochem*, 126: 359-65.

Qi, E., D. Wang, Y. Li, G. Li, and Z. Su. 2019. 'Revealing favorable and unfavorable residues in cooperative positions in protease cleavage sites', *Biochem Biophys Res Commun*, 519: 714-20.

Rasmussen, H. B., S. Branner, F. C. Wiberg, and N. Wagtmann. 2003. 'Crystal structure of human dipeptidyl peptidase IV/CD26 in complex with a substrate analog', *Nat Struct Biol*, 10: 19-25.

Rawlings, N. D., A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn. 2018. 'The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database', *Nucleic Acids Res*, 46: D624-D32.

Rosenthal, P. B., and R. Henderson. 2003. 'Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy', *J Mol Biol*, 333: 721-45.

Ross, B., S. Krapp, M. Augustin, R. Kierfersauer, M. Arciniega, R. Geiss-Friedlander, and R. Huber. 2018. 'Structures and mechanism of dipeptidyl peptidases 8 and 9, important players in cellular homeostasis and cancer', *Proc Natl Acad Sci U S A*, 115: E1437-E45.

Schechter, I., and A. Berger. 1967. 'On the size of the active site in proteases. I. Papain', *Biochem Biophys Res Commun*, 27: 157-62.

Schenk, A. D., S. Cavadini, N. H. Thoma, and C. Genoud. 2020. 'Live Analysis and Reconstruction of Single-Particle Cryo-Electron Microscopy Data with CryoFLARE', *J Chem Inf Model*, 60: 2561-69.

Scheres, S. H. 2012. 'RELION: implementation of a Bayesian approach to cryo-EM structure determination', *J Struct Biol*, 180: 519-30.

Soneson, Charlotte, Vytautas Iesmantavicius, Daniel Hess, Michael Stadler, and Jan Seebacher. 2023. 'einprot: flexible, easy-to-use, reproducible workflows for statistical analysis of quantitative proteomics data', *Journal of Open Source Software*, 8: 5750.

Thoma, R., B. Loffler, M. Stihle, W. Huber, A. Ruf, and M. Hennig. 2003. 'Structural basis of proline-specific exopeptidase activity as observed in human dipeptidyl peptidase-IV', *Structure*, 11: 947-59.

Vizovisek, M., R. Vidmar, M. Fonovic, and B. Turk. 2016. 'Current trends and challenges in proteomic identification of protease substrates', *Biochimie*, 122: 77-87.

Walter, R., W. H. Simmons, and T. Yoshimoto. 1980. 'Proline specific endo- and exopeptidases', *Mol Cell Biochem*, 30: 111-27.

Wang, L., N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, and C. Fu. 2022. 'Therapeutic peptides: current applications and future directions', *Signal Transduct Target Ther*, 7: 48.

Wang, R. Y., Y. Song, B. A. Barad, Y. Cheng, J. S. Fraser, and F. DiMaio. 2016. 'Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta', *Elife*, 5.

Williams, C. J., J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall, 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, and D. C. Richardson. 2018. 'MolProbity: More and better reference data for improved all-atom structure validation', *Protein Sci*, 27: 293-315.

Zhang, K. 2016. 'Gctf: Real-time CTF determination and correction', *J Struct Biol*, 193: 1-12.

Zhang, X., M. J. Belousoff, Y. L. Liang, R. Danev, P. M. Sexton, and D. Wootten. 2021. 'Structure and dynamics of semaglutide- and taspoglutide-bound GLP-1R-Gs complexes', *Cell Rep*, 36: 109374.

Zhang, X., M. J. Belousoff, P. Zhao, A. J. Kooistra, T. T. Truong, S. Y. Ang, C. R. Underwood, T. Egebjerg, P. Senel, G. D. Stewart, Y. L. Liang, A. Glukhova, H. Venugopal, A. Christopoulos, S. G. B. Furness, L. J. Miller, S. Reedtz-Runge, C. J. Langmead, D. E. Gloriam, R. Danev, P. M. Sexton, and D. Wootten. 2020. 'Differential GLP-1R Binding and Activation by Peptide and Non-peptide Agonists', *Mol Cell*, 80: 485-500 e7.

Zhang, Y., B. Sun, D. Feng, H. Hu, M. Chu, Q. Qu, J. T. Tarrasch, S. Li, T. Sun Kobilka, B. K. Kobilka, and G. Skiniotis. 2017. 'Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein', *Nature*, 546: 248-53.

Zivanov, J., T. Nakane, B. O. Forsberg, D. Kimanius, W. J. Hagen, E. Lindahl, and S. H. Scheres. 2018. 'New tools for automated high-resolution cryo-EM structure determination in RELION-3', *Elife*, 7.